Department of Statistics

COURSE STATS 330

Model answer to Assignment 4, 2005

1. *Read the data into R and make a data frame ICU.df.*

The data are in the file ICU.txt. Copy it to a suitable directory and make the data frame:

```
ICU.df<-read.table(file.choose(), header=T)
```

Note that several of the variables are categorical, most have two levels coded 0/1 so are already in the form of dummy variables. However, the variables RACE and LOC each have 3 levels so must be turned into factors:

```
FLOC<-factor(ICU.df$LOC)
FRACE<-factor(ICU.df$RACE)
ICU.df<-data.frame(ICU.df[,-c(5,21)],LOC=FLOC, RACE=FRACE)
```

[5 marks]

2. *Using all the variables, fit a logistic regression model to the data which explains the probability of death in terms of the other variables. Using subset selection, or otherwise, identify those variables that seem to have an influence on ICU mortality. Refit the model using these variables, and carefully discuss how they influence mortality. Do any of the explanatory variables in the reduced model require transformation?*

Using all the variables, we fit the model:

```
ICU.glm<- glm(STA ~ AGE  + SEX + RACE + SER + CAN + CRN
+ INF + CPR + SYS + HRA + PRE + TYP + FRA + PO2 + PH + PCO +
BIC + CRE + LOC, family=binomial,data=ICU.df)
```

[5 marks] Deduct 3 marks if the variables LOC and RACE are not turned into factors.

Next set up the null model for a stepwise regression:

```
null.glm<-glm(STA ~ 1, family=binomial,data=ICU.df)
```

and do the stepwise regression

```
step(null.glm, scope=formula(ICU.glm), direction="both")
```

This results in the model

```
STA ~ LOC + TYP + SYS + CAN + AGE
```

[5 marks] for stepwise. Don't deduct marks if no anova is done.

To check if the continuous variables need to be transformed, we fit a gam model

```
library(mgcv)
plot(gam(STA ~ LOC + TYP +s(SYS) + CAN + s(AGE), family =
binomial, data = ICU.df))
```

The plots show straight lines indicating that SYS and AGE don't need to be transformed. Fitting polynomials also indicates the same thing.

[5 marks] Give 2 marks for gam plot and 3 marks for a correct conclusion

The selected model is fitted:

```
ICUsub.glm<- glm(STA ~ LOC + TYP + SYS + CAN + AGE, family =
binomial,  data = ICU.df)

> summary(ICUsub.glm)

Call:
glm(formula = STA ~ LOC + TYP + SYS + CAN + AGE, family =
binomial,
    data = ICU.df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.16314  -0.54198  -0.31837  -0.07933   2.63978

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.433e+00  1.792e+00  -3.031  0.00243 **
LOC1         2.154e+01  1.457e+03   0.015  0.98820
LOC2         2.417e+00  8.743e-01   2.764  0.00570 **
TYP          3.848e+00  1.270e+00   3.030  0.00244 **
SYS         -1.780e-02  7.473e-03  -2.382  0.01722 *
CAN          2.597e+00  9.626e-01   2.698  0.00697 **
AGE          3.832e-02  1.294e-02   2.962  0.00306 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 128.44  on 193  degrees of freedom
AIC: 142.44

Number of Fisher Scoring iterations: 16
```

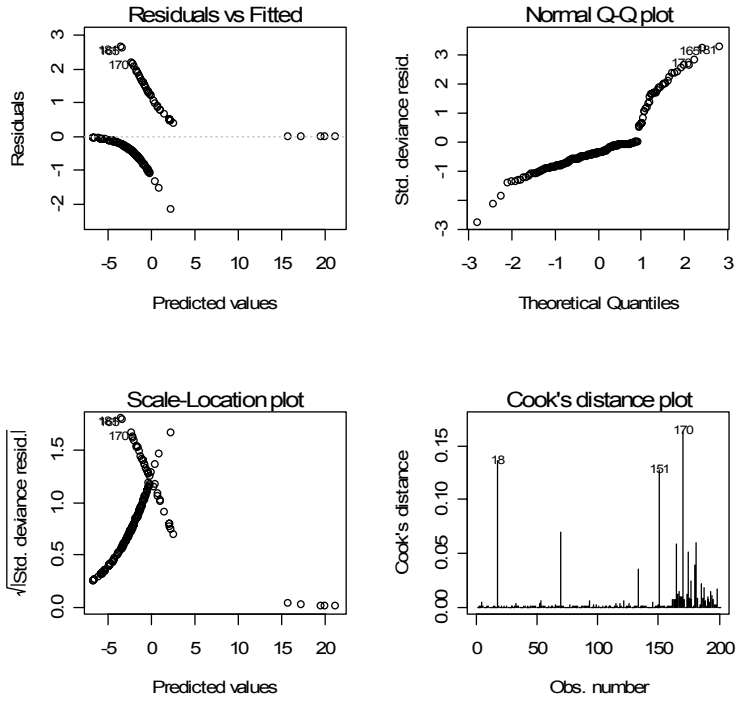
The effect of the variables is as follows:
LOC: Both deep stupor and coma increase the chance of mortality, stupor is more serious than coma,
TYP: There is a higher chance of death from an emergency admission
SYS: Low systolic BP decreases the chance of survival,
CAN: Cancer patients have a lower chance of survival,
AGE: Older patients have a lower chance of survival.

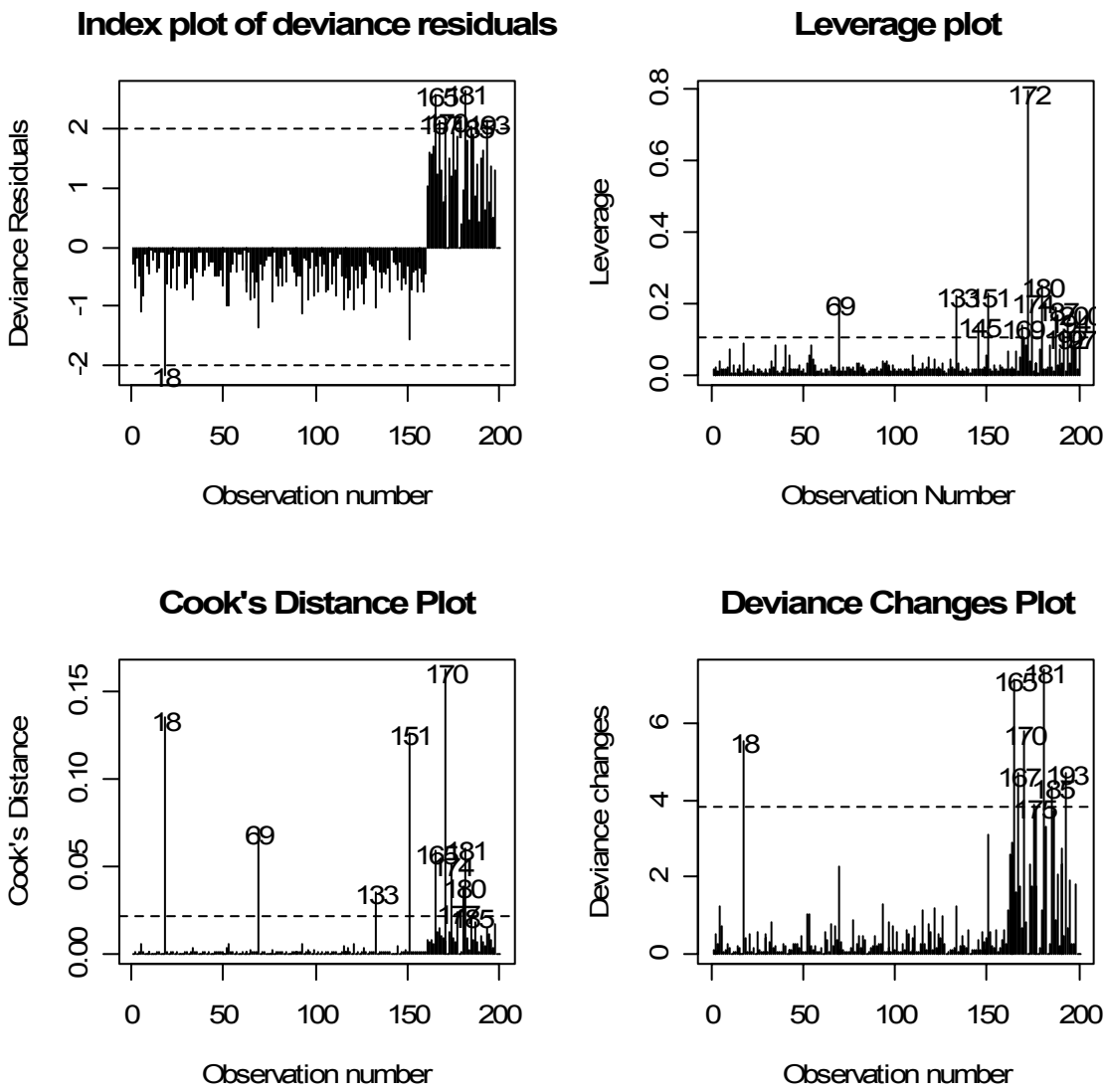
[5 marks] *2 for fitting model and 3 for correct interpretation*

3. *Subject your model to the usual diagnostic checks. Are there any points that have an influence on the fitted model? What points are they?*

NOTE: For the reasons mentioned in class, it is not necessarily a good idea to delete influential points in logistic regression. In this case, retain them.

The functions `plot` and `gam.diag.plots` are useful to identify influential points.





From these plots we see several interesting things. First, there are several points that have very large fitted log-odds, and very small residuals. We can identify these by typing

```
sort(predict(ICUsub.glm ))
```

and reading off the case numbers. They are cases 172, 178, 183, 199 and 200, and correspond to the cases with LOC=1.

All these cases died. The model is fitting these points well, but they may be distorting the results. To check this, we deleted the 5 points with LOC=1 from the regression and refitted. There was no change, so we retained the points.

Next, we identify the most influential points as measured by Cook's distance. We use the code

```
sort(cooks.distance(ICUsub.glm))
```

to identify the points. The 5 points having the biggest Cook's distances are

181 69 151 18 170

These points are influential on the following criteria:

	Response	Est prob	Deviance Res	Cooks D	Hat value
181	1	0.031	2.640	0.060	0.013
69	0	0.605	-1.363	0.070	0.203
151	0	0.703	-1.558	0.127	0.226
18	0	0.904	-2.163	0.136	0.085
170	1	0.093	2.179	0.163	0.096

We refitted the model deleting each of these points in turn and also deleting all 5. The following table shows the changes in the coefficients:

	None	181	69	151	18	170	All 5
(Intercept)	-5.43	-5.91	-5.95	-6.12	-4.90	-20.45	-22.61
LOC1	21.54	21.77	22.00	21.38	21.82	38.25	41.04
LOC2	2.42	2.47	2.41	3.19	3.29	2.41	21.83
TYP	3.85	3.93	4.45	3.87	3.88	18.99	20.68
SYS	-0.02	-0.02	-0.02	-0.01	-0.02	-0.02	-0.02
CAN	2.60	2.79	3.31	2.59	2.67	2.37	3.42
AGE	0.04	0.05	0.04	0.04	0.04	0.04	0.05

The results of deleting all 5 are rather startling. Typically, deleting influential points will increase the size of the coefficients, often to a large extent. Should we delete point 170? It is having a very big effect on the fit.

If we changed the response of point 70 from a 1 to a zero, what would happen? We get

	170=1	170=0
(Intercept)	-5.433	-20.516
LOC1	21.544	38.318
LOC2	2.417	2.406
TYP	3.848	19.059
SYS	-0.018	-0.018
CAN	2.597	2.366
AGE	0.038	0.036

Thus, the coefficients are very much dependent on the value of the response of point 170.

Should we change it? Probably not, as the model predicts a 1 with about a 10% chance. This is not such an unlikely event. However, we should be aware that the analysis is sensitive to this point.

[8 marks] 4 marks for the two plots, and 4 marks for correctly identifying the influential points.

4. *The purpose of the study was to develop a "prognostic index" that would allow the clinicians to predict if an admitted patient will survive or not. A patient is predicted to die if the estimated probability of death exceeds 0.5. How good is your model as a predictor? A claim is made that the predictor (when used as a classifier on new cases) has the following properties*

   Sensitivity: 42%
   Specificity: 95%
   Percentage correctly classified: 84%

   Do you agree with this claim? Hint: use 10-fold cross-validation. I have added a piece of sample code to do cross-validation overleaf. You may need to modify it slightly.

We modify the program given as follows:

```
n<-length(residuals(ICU.glm)) # n is number of cases

nfold<- 10

# divide data into 10 subsamples, each of size m
m<-n%/%nfold

# make vectors to store sensitivity, specificity and total
misclssified for each sub-sample
Sense<-numeric(nfold)
Spec<-numeric(nfold)
Total<- numeric(nfold)


# put in a formula to describe your model e.g

formula<- STA ~ LOC + TYP + SYS + CAN + AGE

rand.order<-sample(n)
ICU.df<-ICU.df[rand.order,]

# this randomly rearranges the rows of the data frame
```

```
sample<-1:m  # first pick the first m rows

# now for a loop using each of the subsamples in turn

for(j in 1:nfold){
# fit model using data not in subsample
fit<-glm(formula, family = binomial, data = ICU.df[-sample,])
# use model to predict sample y-values
y.pred<-(predict(fit, ICU.df[sample,],type="response")>0.5)*1
#get actual sample y-values
y.act<-ICU.df$STA[sample]
#calculate proportion of correctly predicted deaths
Sense[j]<-sum(y.act*y.pred)/sum(y.act)

#calculate proportion of correctly predicted survivors
Spec[j]<- sum((1-y.act)*(1-y.pred))/sum(1-y.act)

#calculate proportion of correct predictions
Total[j]<- ( sum(y.act*y.pred) +  sum((1-y.act)*
 (1-y.pred)))/length(y.pred)


# get new sub-sample
sample<-sample + m
}
# average the nfold results
Av.Sense<- mean(Sense, na.rm=T)
Av.Spec<- mean(Spec, na.rm=T)
Av.Total<- mean(Total, na.rm=T)

Running this gives
> Av.Sense
[1] 0.4195238
> Av.Spec
[1] 0.9717105
> Av.Total
[1] 0.855
```

which agrees pretty well with the claim.

[7 marks]. Note that here we are identifying a death with a "success". Deduct 3 marks if the interpretation is reversed i.e. if a survival is a 'success". (This is contrary to the interpretation given on the homework sheet.)